Tufts University

EM-0212: Applied Data Science
05-09-2019

Data Project Final Report: Real Estate in Coastal Maine

Shea T. Nelson
shea.nelson@tufts.edu

I. Purpose

The purpose of this document is to provide a high-level summary of a data analysis project in *EM-0212: Applied Data Science*. The analysis outlined herein will characterize the effort to analyze home ownership demographics and the efficacy of local government property tax initiatives with respect to rising out-of-state real estate investment in coastal Maine towns.

II. Scope

The scope of this analysis is limited to records of real estate ownership, sale, and taxation in the rural town of Harpswell, Maine. The data is available in electronic form from 2014-2018 and will form the basis of this investigation. In addition to providing a sense of the current situation, a predictive model to examine the effect of potential policy changes will be developed. Additionally, the effect of tax changes on overall tax base will be investigated.

III. Background

The town of Harpswell is a small, rural community on the coast of Maine. For most of its history, Harpswell has depended upon farming and fishing to support its economy. As such, many families have large generational properties situated directly on the coast. However, over the last several decades, the town has become a popular vacation destination for many wealthy families in New England, the rest of the United States, and even abroad. This has caused a dramatic increase in property values, with some areas fetching close to $1 million per acre. Families who had lived on the same land for centuries soon found themselves unable to keep up with tax payments and began to move away. As such, the town faced a demographic and political crisis and was under immense pressure to devise a solution. This came in the form of lowering the mill rate to approximately $5.00[*]. Although this stymied the outflow of local families, the efficacy of the policy change remains understudied. Indeed, it is unknown whether the trend has been reversed or merely slowed, and there are not any means to investigate the effects of further action available to the town authorities.

IV. Data Utilized:

The data to be studied is available here: Harpswell Town Assessor's Office

The dataset that forms the basis of this investigation is the 2014-2018 property tax commitment records and 2014-2018 property sales records available online from the Harpswell Tax Assessor's Office. Each year is available as a separate Excel spreadsheet and downloadable for free.

The data itself is focused on the sale, ownership, taxation, and valuation of real estate within the town. The sale date, land value, structure value, and extensive owner information are available on nearly every plot in the town. This will provide a way to thoroughly examine the current state of affairs as well as identify larger trends to utilize for predictions. Though it is extremely comprehensive and granular, the data does have a number of formatting issues and some structural inconsistencies. This is not insurmountable, however, and cleaning the data will almost certainly yield very much useful datasets. The statistical analysis software JMP will be

[*] In this case, "mill rate" refers to the dollar value in tax to be paid, per thousand dollars of real estate value.

required in order to accomplish this, however, as the data set will contain several thousand lines once fully assembled.

For a full description of all variables and more detailed analysis of potential issues with the data, see Appendix A and Appendix B.

Investigator's Note: See Attached File "*NELSON_Datasheet.pdf*" for a detailed description of data provenance and important considerations with regards to its current and future usage.

V. Questions

There are two major questions that the investigator will attempt to answer through the analysis of the aforementioned dataset.

1. *Is real estate value a significant factor and indicator of residency status?*
2. *If so, do nonresidents enjoy higher home values?*

The nature of this investigation requires the first question to be addressed before moving on to the second.

VI. Potential Sponsors

The Harpswell town government is a major potential sponsor of this work as it has a vested interest in being able to gauge the results of the mil rate reduction. Indeed, the reduction in the property tax reduced the tax base and thus being able to verify that this was not for nothing is very important. The creation of a prediction element to the study is especially relevant to town proceedings because it would both generate continued solutions to the problem and also allow town policymakers to have a sense of what policy changes could do to the demographics and tax income of the town. This project has the potential to provide options to the town government to achieve their demographic goals while maintaining, or even expanding, the tax base. This would provide a direct monetary benefit to the town as a whole.

Additionally, there are several groups that represent local families, especially those in the fishing industry. These organizations work to preserve and protect the way of life of their members, including their property. This report could prove a valuable asset to local groups when lobbying for change at the local government level. A data-analysis take on the problem would allow for data-driven policy to be suggested and perhaps taken more seriously.

VII. Project Outline

This project has been organized utilizing the Data-Outputs-Actions Framework. The chart below (Fig. 1) outlines the process that will be undertaken in order to answer the three fundamental questions that define the project aim.

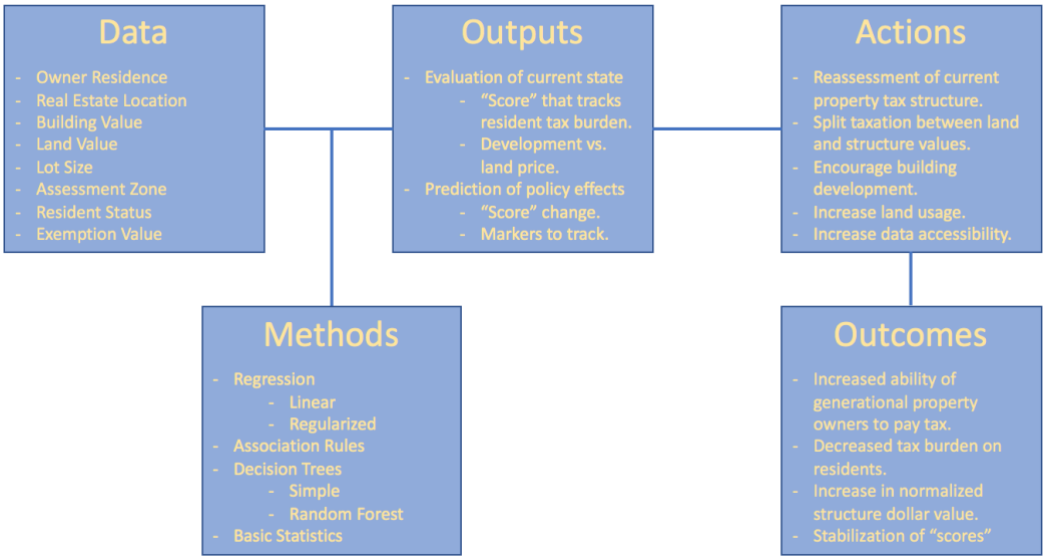

Fig. 1: *Data-Outputs-Action Framework for Harpswell Real Estate Project.*

VIII. Literature Review

No publicly available studies directly involving the dataset to be analyzed in this project have been conducted to the knowledge of the investigator. On the town website, (Harpswell Town Website) no listings or information about what is done with the data are present, although the data itself is freely available. Additionally, no known investigations into the effect of mill rate on local resident ownership rates have been published.

However, despite the lack of scholarship on this particular application, a large body of research exists relating to the economic and social effects of property taxes and mill rates. Most investigations focus on the capitalization of property tax and its effect on the housing market (e.g. Palmon and Smith (1998), Passour (1973), Sirmans, Gatzlaff, and Macpherson (2008), and McDonald (1993)). Within this realm, researchers have focused on several distinct microcosms in order to extrapolate generally about the housing market. Indeed, in Passour (1973) the North Carolinian Farm Market is examined and extrapolated while in McDonald (1993) it is instead Downtown Chicago that is used to make broad statements about the real estate market. This is a point where the scope of this study differs greatly from previous literature in that it is intended to answer a small question for a small town and does not intend to extrapolate the results to trends *larger than the area of study itself*.

Another large area of study is the effect of property tax on equality. Most investigations mention it in some capacity and there are several that address equity in property tax directly. Indeed, in Sunderman et al. (1990) the full range of equity models are evaluated with respect to the Chicago suburb housing market. This is essentially the closest most other studies have gotten to the subject of this proposal which is examining the effect of property tax on local residential ownership. Other papers, such as Sirmans et al. (2008) and Palmon and Smith (1998), also address aspects that deal with the effect on the homeowner but do not address the core focus of this proposal either.

In conclusion, although many studies have been conducted dealing with other aspects and implications of property taxes, none have addressed local vs. nonlocal ownership rate specifically and certainly not in the context of rural Maine.

See Appendix D for Annotated Bibliography

IX. Potential Roadblocks

Though there is a large amount of data available within the dataset in question, there are several potential roadblocks. Firstly, there is simply a lot of data. Though the computer programs to be utilized can easily handle it, it is difficult as a human to keep track of and visualize so much data. Secondly, there are various formatting issues with the data. A common example is that different abbreviations are used for addresses (e.g. road vs. RD). This will make selecting large groups of data difficult as character-based selections will not grab the entirety of relevant data. However, cleaning the data to standardize abbreviations will likely help abate this problem. Another formatting problem is that some numerical entries, such as with land area and value, are listed as 0. This is a nonsensical value for something that must have area and would become an issue in, for example, standardizing with respect to land because it would result in a division by 0. This can also be addressed in data cleaning by replacing it with a value that simulates 0 or by removing it entirely. Also important is the fact that the data comes from several different files. This will require combining the files together to form one large data set. The issues highlighted earlier will make this more difficult, as will inconsistencies between the datasets. For example, the variable that lists street names is titled “Address 1” in one file and “Physical Location” in another file. This will have to be addressed when cleaning the data.

For more information on the data, including a description of all variables and their potential issues, see Appendix A and Appendix B.

X. Process

This investigation was conducted by the construction of a Jupyter Labs Notebook. The overall structure of the investigation was to integrate the tax data with the real estate sales data in order to create a good picture of the current situation. Then, Random Forest Analysis was used to verify that real estate value was significantly important with regards to residency. Then, the trend over the last five years was investigated by integrating real estate sales data over the previous half-decade into a trend dataset. This dataset was used to quantify the current trend, as explore changes to the trend over time. As with the previous dataset, random forest analysis was conducted to verify the importance of real estate value with regards to residency.

XI. Results

The data analysis showed that, in general, nonresidents have higher real estate values than residents. Fig. 2 shows that all quartiles of nonresidents are greater than their equivalents for residents. Real estate value is also shown to be a reasonably powerful predictor of residency in Fig. 3. Compared to the trend box plot in Fig. 4, the difference between residents and nonresidents in the merged dataset seems small. However, the trend dataset deals with only sales over the last four years and reveals the trend that the character of ownership in the town

is changing. However, this trend appears to be slackening. Despite that, real estate value is shown, in Fig. 5, to be a powerful predictor of residency in terms of purchases.

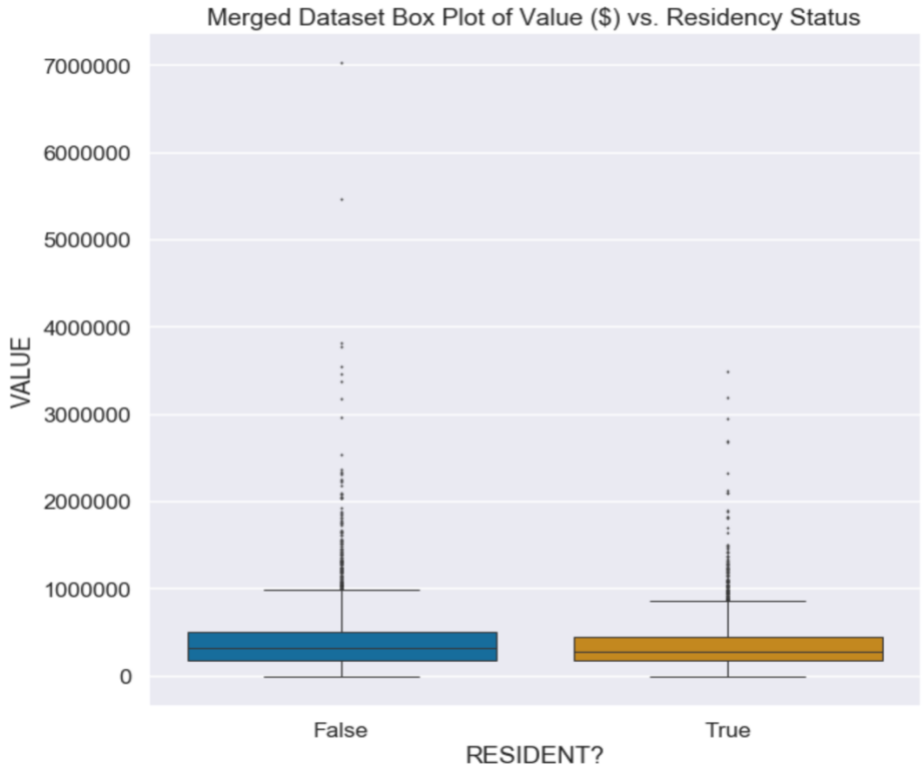

Fig 2: Box Plot for Value vs. Residency, Merged Dataset

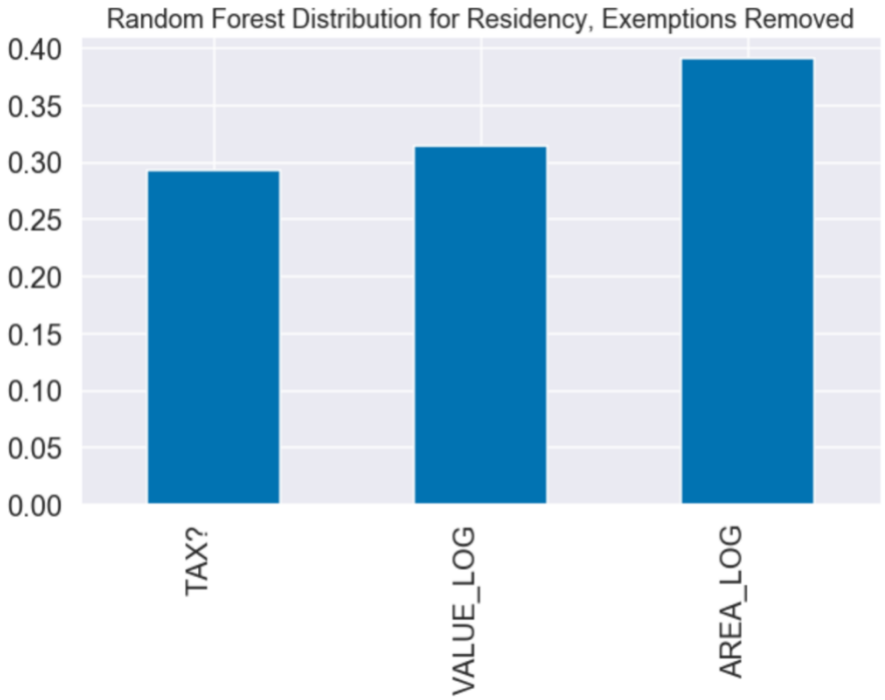

Fig. 3: Random Forest Results for Merged Dataset

Fig. 4: Box Plot for Value vs. Residency, Trend Dataset.

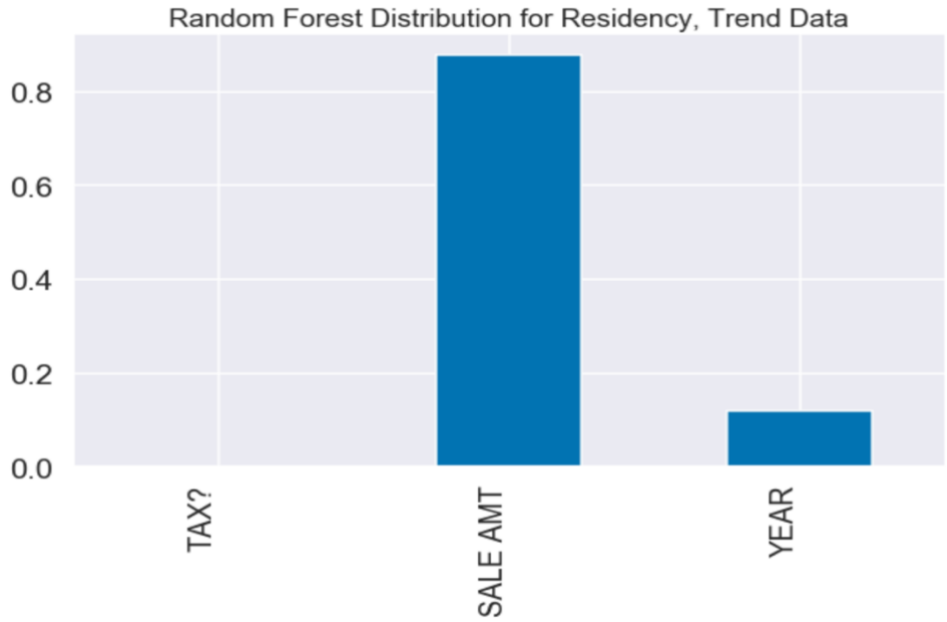

Fig. 5: Random Forest Results for Trend Dataset

XII. Conclusions

The data indicates that the average real estate value for nonresidents is significantly higher than that of residents. The first two random forest analysis directly link real estate valuation with prediction of residency. Based on the previous results it indicates that valuation of real estate is an indicator of residency with nonresidents enjoying significantly higher values thereof. Thus, it can be concluded that nonresidents tend to not only own real estate that is significantly more valuable than residents, but that they contribute directly to a trend of increasing sale prices and increased values across the board. However, it should be noted that the predictive power of the random forest model linking value to residency is not particularly good and the findings require further investigation.

In terms of trends, it is confirmed that there is a gap between value of real estate purchased by nonresidents vs. residents, at least over the past five years. The value of the real estate purchased by nonresidents is significantly greater across all quartiles and has extremely high outliers. However, since the 2018 data has a smaller gap between the residents and nonresidents compared to the 5-year data, there is some indication that the trend is reversing or at least being mitigated. Additionally, the random forest analysis confirmed that Sales Amount was by far the greatest predictor of residency status. This narrows down the possibilities of significant factors to explain the perceived disparity in wealth between residents and nonresidents.

XIII. Reflections

The investigator for this project had absolutely no experience with Python beforehand. Thus, he relied heavily upon the lectures and notes provided by Prof. Forde cited in the document. In the future it would be a good idea to better evaluate the power of the random forest analysis and perhaps investigate other correlation prediction models such as regression analysis. Indeed, there were not enough variables under study with the random forest analysis to use it to its full potential. This was only realized by the investigator after the fact and thus remains an open area of study. The data may be over processed as the investigator feels that he made too many efforts to clean the data beyond making it simply useful. This may have inadvertently eliminated potentially important sources of data for analysis. Furthermore, data is available as far back as 2014 for the town overall which would help tease out the trends in greater detail. The trends themselves require more study as the fact that each year is delineated was an avenue of investigation was under-analyzed by the investigator. Another important area would be attempting to reconcile the results of this and future investigations with town policies enacted. Additionally, the town has further data available offline which could be incorporated to better understand the situation.

Appendix

Appendix A: *List of Variables and Associated Descriptions, Tax Data*

Variable	Type	Description	Identified Issues
RealEstate_ID	Numeric	Real estate identification number, unique.	Several numbers have characters mixed in and are of completely different format.
Owner_Name	Character	First entry for owner of real estate.	Formatting of name is variable and not consistent. Some people are repeats but own multiple properties.
Owner_Name2	Character	Second entry for owner of real estate.	Formatting of name is variable and not consistent. Some people are repeats but own multiple properties. Also includes random elements that do not fit in first entry.
Address1	Character	Street address of real estate owner's primary residence.	Names of trusts and other organizations often replace street address.
Address2	Character	Second entry for street address of real estate owner's primary residence.	Includes mix of modifiers (e.g. APT #2) and displaced street addresses.
City	Character	City of real estate owner's primary residence.	Dash is used when owner lives in unincorporated town.
State	Character	State abbreviation of real estate owner's primary residence.	Dash is used when owner is not from USA.
Zip	Numeric	Zip code of real estate owner's primary address.	Some zip codes have 4-digit suffix.
AccountNumber	Character	Real estate identification account number.	NONE
Land_Value	Numeric	Dollar value of land associated with assessed real estate. No units.	Presence of 0 values for land value prevents division by lot value without cleaning data.
Building_Value	Numeric	Dollar value of buildings associated with assessed real estate. No units	Lots of zero values due to land without improvements.
Exemption1	Numeric	Dollar value of first applied exemption, no units.	NONE

Exemption_Code1	Numeric	First entry for tax exemption code.	NONE
Exemption2	Numeric	Dollar value of second applied exemption, no units.	Barely any nonzero values.
Exemption_Code2	Numeric	Second entry for tax exemption code.	Barely any nonzero values.
Exemption3	Numeric	Dollar value of third applied exemption, no units.	Barely any nonzero values.
Exemption_Code3	Numeric	Third entry for tax exemption code.	Barely any nonzero values.
Total Assessed	Numeric	Total value of real estate assessed in current year, 2018.	Lots of 0 values due to exemptions.
2018 Taxes	Numeric: Currency	Lists total assessed taxes owed for current year, 2018, in dollars.	Lots of $0.00 values due to exemptions.
Resident	Character	Binary character (Y or N) if primary real estate owner is resident of town.	N value applies to individuals as well as organizations.
Property_Location	Character	Lists house number and road of real estate location.	Some addresses lack a numerical prefix. Mixes numerals and characters. Island locations are listed as a road.
Zone	Character	Lists assessment zone.	Some land straddles zones and data is formatted XX/XX instead of just XX.
Lot_Size	Numeric	Lists lot size of assessed real estate in acres.	Presence of 0 values for land area prevents division by lot size without cleaning data.

Appendix B: *List of Variables and Associated Descriptions, Sales Data*

Variable	Type	Description	Identified Issues
Map	Numeric	Tax map identification number.	NONE
Lot	Numeric	Lot identification number.	Lot numbers repeat as each tax map resets lot counter.
Sub	Numeric	Subdivision of lot, number.	Sub numbers repeat as each lot resets sub counter.
BK/PG	Numeric	Book/Page of deed registry.	NONE
Last Name/LLC	Character	Last name or LLC name of new owner.	Personal names are mixed together with business names.
First Name and Others/Trust	Character	First name or other owners or trusts of new owner.	Personal names are mixed together with business names.
New Owner Address	Character	Street and house number of new owner.	Abbreviations and full word are mixed (e.g. RD vs. Road)
City	Character	City of new owner's primary residence.	NONE
ST	Character	State abbreviation of new owner.	If new owner is from out of US, state abbreviation is blank.
Zip	Numeric	Zip code of new owner.	If new owner is from out of US, zip code is blank.
Former Owner	Character	Full name of former owner.	Personal names are mixed together with business names.
Date Signed	Numeric	Date of selling agreement	NONE
Date Record	Numeric	Date of selling record	NONE
Sales AMT	Numeric	Sale price of real estate.	Lots of zero values.
Physical #	Numeric	Number of real estate address.	Several missing values.
Physical Location	Character	Location/Road of real estate address.	Abbreviations and full word are mixed (e.g. RD vs. Road)

Appendix C: *Project Data Source*

http://www.harpswell.maine.gov/?SEC=B8009038-83C1-4082-8297-FFB27C3B4AF0
Note: 2014-2018 Tax commitment and Lists of Property Sales.

Appendix D: *Annotated Bibliography*

1. McDonald, John F. "INCIDENCE OF THE PROPERTY TAX ON COMMERCIAL REAL ESTATE: THE CASE OF DOWNTOWN CHICAGO." *National Tax Journal* 46, no. 2 (June 1993): 109-20. Accessed March 5, 2019. https://www.jstor.org/stable/41789004?seq=1#page_scan_tab_contents.

An empirical study of office rents and property taxes in downtown Chicago from the early 1990's. Shows that 45% of tax differentials are shifted forward to tenants in the form of higher rents. Also shows that values per square foot stems from rent per square foot.

2. Palmon, Oded, and Barton A. Smith. "New Evidence on Property Tax Capitalization." *Journal of Political Economy* 106, no. 5 (1998): 1099-111. Accessed March 5, 2019. https://www.journals.uchicago.edu/doi/pdfplus/10.1086/250041.

Empirical analysis of tax capitalization across 50 subdivisions with similar demographics and amenities in the Northern suburbs of Houston, Texas. Tax capitalization was estimated and compared to the Tiebout Hypothesis about full capitalization. The study implies that the individuals in the housing market do rationally discount properties with higher tax rates and only unexpected tax changes are passed on to new buyers without capitalization. From 1998

3. Pasour, E. C., Jr. "Real Property Taxes and Farm Real Estate Values: Incidence and Implications." *American Journal of Agricultural Economics* 55, no. 4 (November 1, 1973): 549-56. Accessed March 5, 2019. https://academic.oup.com/ajae/article-abstract/55/4_Part_1/549/156025.

Examines the relationship between reductions in property taxes and farm real estate value. In North Carolina. Results show property tax reductions for farm real estate are capitalized into higher real estate values. From 1973. Also shows that farms and large swaths of land benefit less from reduction in property tax because houses don't capitalize as much

4. Siniavskaia, Natalia, PhD. "PROPERTY TAX RATES AFTER THE HOUSING DOWNTURN." *National Association of Home Builders*, April 4, 2011. Accessed March 5, 2019. http://www.nahbclassic.org/generic.aspx?sectionID=734&genericContentID=155396&channelID=311. Report available to the public as a courtesy of HousingEconomics.com

Relies on real estate data from all metro and sub-metro areas in the United States and organizes within state-level context. Discusses the huge decrease in home values following the Great Recession and the inability for assessor adjustments to keep up. This caused an enormous spike in effective tax rate even though mill and property rates remained largely the

same. Essentially, when home values rapidly increase, real tax rate lags and decreases. However, when home values rapidly decrease, real tax rate lags and increases.

5. Sirmans, Stacy, Dean Gatzlaff, and David Macpherson. "The History of Property Tax Capitalization in Real Estate." *Journal of Real Estate Literature* 16, no. 3 (2008): 327-44. Accessed March 5, 2019. https://www.aresjournals.org/doi/abs/10.5555/reli.16.3.557w0080931386u6.

Study on capitalization of property taxes in real estate. Suggests degree of capitalization depends on elasticity of supply. In this case, this means that an increase in demand with inelastic supply will accordingly raise housing prices while an elastic supply means that a change in demand will not alter prices. Also implies that tax capitalization can make it more difficult for homeowners to move. Most studies suggest partial capitalization occurs. From 2008 and essentially a study of previous inquiries into property tax capitalization in real estate.

6. Sunderman, Mark, John Birch, Roger Cannaday, and Thomas Hamilton. "Testing for Vertical Inequity in Property Tax Systems." *Journal of Real Estate Research* 5, no. 3 (1990): 319-34. Accessed March 5, 2019. https://www.aresjournals.org/doi/abs/10.5555/rees.5.3.d528387288447203.

Tests vertical inequity models for real estate property tax burden across class in Chicago suburbs. Contains two data sets. One is “contrived” and used to show potential failings in models. The second set is based on empirical analysis of data collected from condominium sales in a Chicago suburb. Shows that existing equity models are ok for simple cases with continuous first derivatives but break down with more complex and extensive tax environments.